Challenges in the Development of Register-Based Population Statistics

By
Cai Jialin and Tan Yeow Lip
Strategic Resource and Population Division
Singapore Department of Statistics

Introduction

The Singapore Department of Statistics (DOS) has increased its use of administrative data in the production of official statistics since the conduct of the Census of Population (Census) 2000.

In the domain of demographic statistics, two core databases drawing data from multiple administrative sources have been enriched continuously. The core databases comprise (1) a population statistical database with basic demographic information of Singapore's population, and (2) a dwelling statistical database on residential dwellings in Singapore.

These two databases enable the compilation of annual data on basic register-based demographic information such as age, gender, ethnic, geographic location and dwelling type starting from year 2000.

This article provides an overview of the considerations, the historical context as well as the ongoing challenges and strategies in the development of Singapore's register-based population and dwelling statistics.

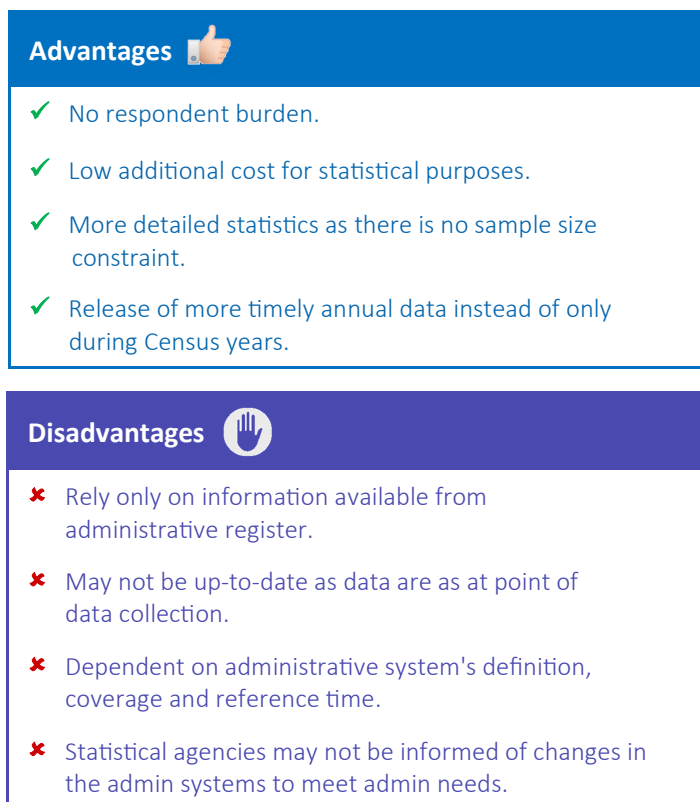## Advantages and Disadvantages of Register-Based Statistics

Register-based statistics is produced by linking up administrative (admin) data from various admin registers.

As admin data are collected to meet the primary functions of admin processes, some units and variables are defined differently from the needs of official statistics.

However, compared to data from surveys, admin data are more readily available, of reasonable quality and less costly. Thus, the use of admin data in place of survey data lowers respondent burden. Figure 1 provides a summary of the advantages and disadvantages of register-based statistics.

Not all data required for the compilation of national statistics are available from the admin registers. For instance, detailed household expenditure items can only be collected through sample surveys. Thus, admin data should not be the only source for official statistics. They complement the data collected through sample surveys.

FIGURE 1   ADVANTAGES AND DISADVANTAGES
OF REGISTER-BASED STATISTICS

Advantages 👍

✔ No respondent burden.

✔ Low additional cost for statistical purposes.

✔ More detailed statistics as there is no sample size constraint.

✔ Release of more timely annual data instead of only during Census years.

Disadvantages ✋

✘ Rely only on information available from administrative register.

✘ May not be up-to-date as data are as at point of data collection.

✘ Dependent on administrative system's definition, coverage and reference time.

✘ Statistical agencies may not be informed of changes in the admin systems to meet admin needs.

The Development of Singapore's Register-Based Statistics

The development of population and dwelling databases was modelled after that of the Nordic countries which had a long history and rich experience in the area of register-based statistics production.

DOS possesses essential pre-conditions that facilitated the extensive use of admin sources in statistics production.

Singapore has an established legal framework to acquire admin data for statistical use, a Unique Identification Number (UIN) and registered address for every resident, and comprehensive and reliable administrative registers developed for admin needs.

Figure 2 details the development of DOS's population and dwelling databases through the years.

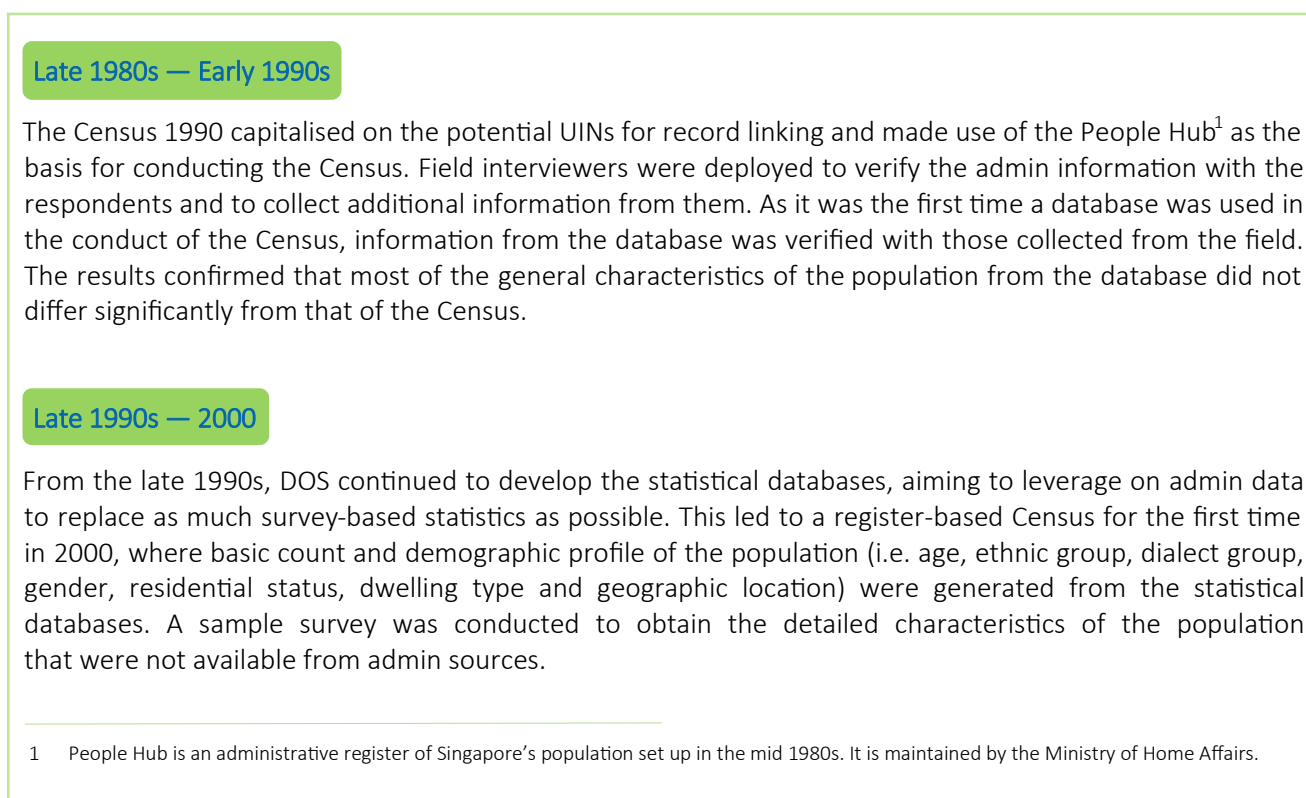FIGURE 2   DEVELOPMENT OF DOS'S POPULATION AND DWELLING DATABASES

Late 1980s — Early 1990s

The Census 1990 capitalised on the potential UINs for record linking and made use of the People Hub[1] as the basis for conducting the Census. Field interviewers were deployed to verify the admin information with the respondents and to collect additional information from them. As it was the first time a database was used in the conduct of the Census, information from the database was verified with those collected from the field. The results confirmed that most of the general characteristics of the population from the database did not differ significantly from that of the Census.

Late 1990s — 2000

From the late 1990s, DOS continued to develop the statistical databases, aiming to leverage on admin data to replace as much survey-based statistics as possible. This led to a register-based Census for the first time in 2000, where basic count and demographic profile of the population (i.e. age, ethnic group, dialect group, gender, residential status, dwelling type and geographic location) were generated from the statistical databases. A sample survey was conducted to obtain the detailed characteristics of the population that were not available from admin sources.

---

1   People Hub is an administrative register of Singapore's population set up in the mid 1980s. It is maintained by the Ministry of Home Affairs.

FIGURE 2   DEVELOPMENT OF DOS'S POPULATION AND DWELLING DATABASES (Cont'd)

Early 2000s — Early 2010s

Broad register-based population estimates and profiles were first released annually in 2001. More register-based data were released to the public via various government platforms for public access such as OneMap[2] and data.gov.sg[3]. After the Census 2010, DOS adopted a four-step iterative process framework to refine and enrich the register-based statistics, starting from concept review to implementation (Figure 3).

2   OneMap is a one-stop integrated geospatial data sharing platform.
3   Data.gov.sg is the Singapore government's one-stop portal to publicly-available datasets.

FIGURE 3   A SYSTEMATIC PROCESS FRAMEWORK FOR REGISTER-BASED STATISTICS

A Four-Step Process Framework
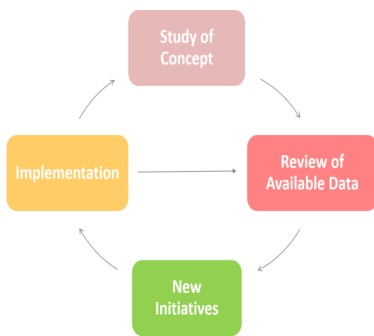
After Census 2010, a systematic four-step process framework was adopted to enhance the data quality of register-based statistics.

Study of Concept

At this fundamental stage, the definition and coverage of new register-based statistics are reviewed by referencing to the recommendations and concepts from the United Nation (UN) and the practices of other advanced National Statistical Offices (NSOs). Some NSOs adopt a different concept for register-based statistics from survey-based statistics.

Review of Available Data

At this stage, the new statistics are constructed with available admin data. Some basic statistics such as the gender and age of a person are compiled from admin data after processing. However, this is not the case for most register-based statistics. Unlike in surveys where questions are formulated in accordance to established definitions, admin data are a by-product of administrative operations. They may thus have divergent definitions, reference periods or coverage since the administrative touch-point may pertain to selected population groups. Therefore, multiple admin sources are cross-referenced to produce more comprehensive and precise register-based statistics.

New Initiatives

New initiatives are conceptualised to resolve the issues identified in the previous step. These include:
(a) exploring additional admin sources;
(b) communicating with data providers to further understand the data content, data quality, data collection and processing procedures;
(c) optimising the use of existing data sources and improving the existing processing rules; and
(d) studying feasible statistical imputation methods.

Implementation

The new initiatives are implemented at this stage and the statistics thus compiled are reviewed. Prior to the release of the statistics, cross-checks against benchmark census data are made. This may be an iterative process when more admin data become available. When the data series has been monitored for a period of time and assessed to be of good quality, it is released for use.

Challenges in Developing Register-Based Statistics

The challenges and approaches taken to transform admin data for statistical uses can be broadly categorised into three aspects, namely: data standardization, data quality and communication with data providers.

Data Standardization

Data standardization transforms the same information collected differently from each admin source into one common standard so that data can be linked.

DOS maintains a well-established code conversion database to convert the various codes from external admin registers into standardized code tables.

In the process of mapping the codes, the two sets of codes must be consistent in terms of definitions and coverage. Figure 4 provides an example on marriages and divorces data.

As a national statistical coordinator, DOS promotes the adoption and use of a common standard in the collection, compilation and dissemination of statistics. This facilitates data sharing and ensures consistency and comparability of data.

FIGURE 4   EXAMPLE OF STANDARDIZATION OF CODES FROM DIFFERENT ADMIN REGISTERS

Statistics on Marriages and Divorces

DOS publishes annual data on marriages and divorces, analysing the marriage and dissolution trends over time. As each admin agency (i.e. Registry of Marriages (ROM), Registry of Muslim Marriages (ROMM), Family Justice Court and Syariah Court) adopts different formats and codes for the same data item, DOS has to map to a standardized code to ensure data comparability. The following example illustrates the standardisation and categorisation of the code values from different admin registers.

Example—Mapping Different Codes for the Same Data Item

Code values adopted by the different source agencies

Previous marital status	Register A	DOS's standardized value
Single	S	1 – Single
Married	M	2 – Married
Widow	F	3 – Widowed
Widower	G	3 – Widowed
Divorced	D	4 – Divorced

Previous marital status	Register B	DOS's standardized value
Bachelor	A	1 – Single
Spinster	B	1 – Single
Married	C	2 – Married
Widowed	D	3 – Widowed
Divorced	E	4 – Divorced

Some challenges in standardizing data from multiple sources are:

Data Stored in Free-Text Description Format

Such data are collected without much standardization and control on quality, and often contain spelling errors and abbreviations. It requires much effort to process, classify and code the data.

DOS encourages the admin sources to store the data in numeric or alphabetical codes instead of free-text descriptions, so as to reduce processing time required for future analysis.

Data are Too Vague

An admin register may use less-detailed classification codes to capture the data.

For example, DOS classifies 'polytechnic diploma' as a category based on the *Singapore Standard Educational Classification* whereas an admin register may capture the data as 'Post-secondary and above' which includes polytechnic diploma, junior college, professional qualifications and other diplomas.

DOS has to assess if the data are still useful for analysis.

Code Changes Over Time by Source Agencies

A code conversion database is then used to map the changes accordingly.

Classification Standards or Definitions Change Over Time but are Classified Under the Same Code by Source Agencies

The changes may not be detected easily if DOS is not informed of them.

DOS performs data consistency checks and trend checks to identify the changes and clarify with the source agency before using the data.

Data Quality

The quality of admin data is reviewed in various aspects, covering: (a) adequacy of information, (b) data consistency, (c) timeliness of data and (d) data coverage.

(a) Adequacy of Information

Admin data may still be insufficient for statistical use despite chaining up multiple admin sources and adopting appropriate statistical methods.

For example, only locally registered marriages are captured officially by admin authorities in Singapore. There is no authority to administer the reporting of overseas marriages.

One possible solution to bridge the data gap is to use birth registration data where the parents' marriage information is obtained at the point of birth registration for local births. This would, however, be a retrospective updating of marriage data since births are likely to take place after marriage.

(b) Data Consistency

There will be occasions when data inconsistencies occur during the process of the combination of multiple data sources. DOS prioritizes the different data sources according to data quality, relevance and timeliness.

Extensive quality checks are put in place to ensure that the data are of good quality before use. Quality checks are broadly categorized as follows:

(i) Micro Checks Covering Intra-Record and Inter-Record Checks

Intra-record checks are performed on the record itself such as a UIN algorithm check. Inter-record checks involve verifying against other records such as age gap between a child and parent.

Figure 5 provides an example of data consistency checks for the Singapore Population Estimates.

(ii) Macro Checks from Cross Sectional and Longitudinal Perspective

Trend checks study data of the same reference period across time, while longitudinal checks track the events occurring for an individual across time.

(iii) Benchmarking against Survey-Based Data

This is one of the essential checks if the register-based statistics are meant to replace survey-based statistics. Besides considering the limitations of admin data accuracy, data limitations and precision of survey-based data have to be taken into account in order to have a balanced assessment.

(iv) Benchmarking against Independent Data from Other Sources

Most of the data from independent data sources do not cover the full population. Thus, the usefulness of such comparison may be limited if the robustness of data are to be assessed on the full population.

(c) Timeliness of Data

Timeliness is an important consideration for admin data. Dates of events play an essential role in determining the period of data collected.

They are however not always available. In such cases, the feasibility and extent of the data to be used has to be assessed.

(d) Data Coverage

Specific sets of definitions and coverage are adopted by different admin agencies for their admin functions.

Admin data are processed for statistical purposes, including the adjustment of scope using other related variables and refinement of reference period using event dates.

FIGURE 5   EXAMPLE OF DATA CONSISTENCY CHECKS ACROSS ADMIN SOURCES

Statistics on Singapore Population Estimates

DOS publishes annual data for resident population which comprises Singapore citizens (SCs) and permanent residents (PRs) with valid local addresses and who are not away from Singapore for a continuous period of 12 months or longer. The basic count of the resident population is based on a person's place of usual residence i.e. de jure concept.

For a resident to be included in the population estimate count, his address must be a valid local address. The following example illustrates inconsistencies in the data where a person wrongly reports an invalid address due to typo or inadequate information. To improve the data quality of such erroneous records, cross-checking of the reported address against other administrative registers is performed. By checking against other sources, such errors are identified and corrected.

	Reported Address (Invalid address)	Address in Register A (Valid address)	Corrected Address
Block	561	561	561
Street	Bank Lane	Bank Lane	Bank Lane
Level	17	17	17
Unit Number	2289	2298	2298
6-digit postal code	477229	477229	477229

Communication with Data Providers

The statistical agency often has limited information on data availability, definitions and treatments of the admin data managed by the source agencies. Close collaboration between the statistical agency and source agencies is essential to optimise the use of admin data.

Besides working closely at the initial stage, it is also important to maintain interactions with the source agencies as:

(i) there could be changes to the admin system or the source agency may stop collecting some data variables which the statistical agency may then be alerted to; or,

(ii) there could also be opportunities to include data items required for statistical uses in the admin registers if the source agency undertakes a review of the admin system and accepts inputs from the statistical agency.

In adopting the "Whole-of-Government" approach for better integration and public service delivery across the Government, admin registers have been useful to DOS in generating register-based official statistics.

Conclusion

Future work in developing register-based population statistics involves obtaining and reviewing the suitability of new data sources, particularly Big Data sources.

DOS has also embarked on initiatives to integrate other Big Data with traditional survey data and admin data.

In addition, the success of developing good quality register-based statistics must be supported by all relevant source agencies. With close collaboration among various agencies and by tapping on Big Data, DOS is working to replace more survey-based data with register-based statistics progressively.

References

B Leow and E Koh (2001): Combining Survey and Administrative Data for Singapore's Census of Population 2000, paper presented at the International Statistical Institute World Statistics Congress 2001.

Singapore Department of Statistics (1999): ESCAP Working Group of Statistical Experts, Eleventh Session: Singapore Census of Population, 2000 - The First Register-Based Census.

Singapore Department of Statistics (2002): Singapore Census of Population 2000: Administrative Report.

Singapore Department of Statistics (2011): Singapore Census of Population 2010: Administrative Report.

United Nations Economic Commission for Europe (2007): Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics.

United Nations Economic Commission for Europe (2016): Big Data in Official Statistics.

Wallgren A. and Wallgren B. (2007): Register-Based Statistics - Administrative Data for Statistical Purposes. John Wiley & Sons, Ltd